The Annals of Statistics
2009, Vol. 37, No. 6A, 3697–3714
DOI: 10.1214/08-AOS663
In the Public Domain

ESTIMATING THE GUMBEL SCALE PARAMETER FOR LOCAL ALIGNMENT OF RANDOM SEQUENCES BY IMPORTANCE SAMPLING WITH STOPPING TIMES

By Yonil Park, Sergey Sheetlin and John L. Spouge

National Library of Medicine

The gapped local alignment score of two random sequences follows a Gumbel distribution. If computers could estimate the parameters of the Gumbel distribution within one second, the use of arbitrary alignment scoring schemes could increase the sensitivity of searching biological sequence databases over the web. Accordingly, this article gives a novel equation for the scale parameter of the relevant Gumbel distribution. We speculate that the equation is exact, although present numerical evidence is limited. The equation involves ascending ladder variates in the global alignment of random sequences. In global alignment simulations, the ladder variates yield stopping times specifying random sequence lengths. Because of the random lengths, and because our trial distribution for importance sampling occurs on a different sample space from our target distribution, our study led to a mapping theorem, which led naturally in turn to an efficient dynamic programming algorithm for the importance sampling weights. Numerical studies using several popular alignment scoring schemes then examined the efficiency and accuracy of the resulting simulations.

1. Introduction. Sequence alignment is an indispensable tool in modern molecular biology. As an example, BLAST [2, 3, 18] (the Basic Local Alignment Search Tool, `http://www.ncbi.nlm.nih.gov/BLAST/`), a popular sequence alignment program, receives about 2.89 submissions per second over the Internet. Currently, BLAST users can choose among only 5 standard alignment scoring systems, because BLAST p-values must be precomputed with simulations that take about 2 days for the required p-value accuracies. Moreover, adjustments for unusual amino acid compositions are

Received October 2007; revised June 2008.
AMS 2000 subject classifications. Primary 62M99; secondary 92-08.
Key words and phrases. Gumbel scale parameter estimation, gapped sequence alignment, importance sampling, stopping time, Markov renewal process, Markov additive process.

essential in protein database searches [33], and in that application, computational speed demands that the corresponding p-values be calculated with crude, relatively inaccurate approximations [3]. Accordingly, for more than a decade, much research has been directed at estimating BLAST p-values in real time (i.e., in less than 1 sec) [7, 24, 26, 29], so that BLAST might use arbitrary alignment scoring systems.

Several studies have used importance sampling to estimate the BLAST p-value [7, 9, 26]. To describe importance sampling briefly, let $\mathbb{E}$ denote the expectation for some "target distribution" $\mathbb{P}$, let $\mathbb{Q}$ be any distribution, and consider the equation

$$\mathbb{E}X := \int X(\omega)\, d\mathbb{P}(\omega) = \int X(\omega) \frac{d\mathbb{P}(\omega)}{d\mathbb{Q}(\omega)}\, d\mathbb{Q}(\omega). \tag{1.1}$$

A computer can draw samples ω_i $(i = 1, \ldots, r)$ from the "trial distribution" $\mathbb{Q}$ to estimate the expectation: $\mathbb{E}X \approx r^{-1} \sum_{i=1}^{r} X(\omega_i)[d\mathbb{P}(\omega_i)/d\mathbb{Q}(\omega_i)]$. The name "importance sampling" derives from the fact that the subsets of the sample space where X is large dominate contributions to $\mathbb{E}X$. By focusing sampling on the "important" subsets, judicious choice of the trial distribution $\mathbb{Q}$ can reduce the effort required to estimate $\mathbb{E}X$. In importance sampling, the likelihood ratio $d\mathbb{P}(\omega)/d\mathbb{Q}(\omega)$ is often called the "importance sampling weight" (or simply, the "weight") of the sample ω.

A Monte Carlo technique called "sequential importance sampling" can substantially increase the statistical efficiency of importance sampling by generating samples from $\mathbb{Q}$ incrementally and exploiting the information gained during the increments to guide further increments. Although sequences might seem an especially natural domain for sequential sampling, most simulation studies for BLAST p-values have used sequences of fixed length. In contrast, our study involves sequences of random length.

Here, as in several other importance sampling studies [7, 9, 26, 34], hidden Markov models generate a trial distribution $\mathbb{Q}$ of random *alignments* between two sequences, where the *sequences* have a target distribution $\mathbb{P}$. The other studies gloss over the fact that their trial and target distributions occur on different sample spaces, such as alignments and sequences. The other studies used sequences of fixed lengths, however, where a relatively simple formula for the weight $d\mathbb{P}/d\mathbb{Q}$ pertains. For the sequences of random length in this paper, however, the stopping rules for sequential sampling complicate formulas for $d\mathbb{P}/d\mathbb{Q}$. Accordingly, the Appendix gives a general mapping theorem giving formulas for the weights $d\mathbb{P}/d\mathbb{Q}$ when each sample from $\mathbb{P}$ corresponds to many different samples from $\mathbb{Q}$. (In the present article, e.g., each pair of random sequences corresponds to many possible random alignments.) In addition to the mapping theorem, we also develop several other techniques specifically tailored to speeding the estimation of the BLAST p-value.

The organization of this article follows. Section 2 on background and notation is divided into 4 subsections containing: (1) a friendly introduction to sequence alignment and its notation; (2) a brief self-contained description of the algorithm for calculating global alignment scores; (3) a technical summary of previous research on estimating the BLAST p-value introducing our importance sampling methods; and (4) a heuristic model for random sequence alignment using Markov additive processes. Section 3 on Methods is also divided into 4 subsections containing: (1) a novel formula for the relevant Gumbel scale parameter λ; (2) a Markov chain model for simulating sequence alignments (borrowed directly from a previous study [34], but used here with a stopping time); (3) a dynamic programming algorithm for calculating the importance sampling weights in the presence of a stopping time; and (4) formulas for the simulation errors. Section 4 then gives numerical results for the estimation of λ under 5 popular alignment scoring schemes. Finally, Section 5 is our Discussion.

2. Background and notation.

2.1. *Sequence alignment and its notation.* Let $\mathbf{A} = A_1A_2\cdots$ and $\mathbf{B} = B_1B_2\cdots$ be two semi-infinite sequences drawn from a finite alphabet $\mathfrak{L}$, for example, $\{\mathrm{A},\mathrm{C},\mathrm{D},\mathrm{E},\mathrm{F},\mathrm{G},\mathrm{H},\mathrm{I},\mathrm{K},\mathrm{L},\mathrm{M},\mathrm{N},\mathrm{P},\mathrm{Q},\mathrm{R},\mathrm{S},\mathrm{T},\mathrm{V},\mathrm{W},\mathrm{Y}\}$ (the amino acid alphabet) or $\{\mathrm{A},\mathrm{C},\mathrm{G},\mathrm{T}\}$ (the nucleotide alphabet). Let $s:\mathfrak{L}\times\mathfrak{L}\mapsto\mathbb{R}$ denote a "scoring matrix." In database applications, $s(a,b)$ quantifies the similarity between a and b, for example, the so-called "PAM" (point accepted mutation) and "BLOSUM" (block sum) scoring matrices can quantify evolutionary similarity between two amino acids [11, 16].

The alignment graph $\Gamma_{\mathbf{A},\mathbf{B}}$ of the sequence-pair $(\mathbf{A},\mathbf{B})$ is a directed, weighted lattice graph in two dimensions, as follows. The vertices v of $\Gamma_{\mathbf{A},\mathbf{B}}$ are nonnegative integer points (i,j). (Below, ":=" denotes a definition, e.g., the natural numbers are $\mathbb{N} := \{1,2,3,\ldots\}$. Throughout the article, i,j,k,m,n and g are integers.) Three sets of directed edges e come out of each vertex $v=(i,j)$: northward, northeastward and eastward (see Figure 1). One northeastward edge goes into $v=(i+1,j+1)$ with weight $s[e]=s(A_{i+1},B_{j+1})$. For each $g>0$, one eastward edge goes into $v=(i+g,j)$ and one northward edge goes into $v=(i,j+g)$; both are assigned the same weight $s[e]=-w_g<0$. The deterministic function $w:\mathbb{N}\mapsto(0,\infty]$ is called the "gap penalty." (The value $w_g=\infty$ is explicitly permitted.) This article focuses on affine gap penalties $w_g=\Delta_0+\Delta_1 g$ $(\Delta_0,\Delta_1\geq 0)$, which are typical in BLAST sequence alignments. Together, the scoring matrix $s(a,b)$ and the gap penalty w_g constitute the "alignment parameters."

A (directed) path $\pi=(v_0,e_1,v_1,e_2,\ldots,e_k,v_k)$ in $\Gamma_{\mathbf{A},\mathbf{B}}$ is a finite alternating sequence of vertices and edges that starts and ends with a vertex. For

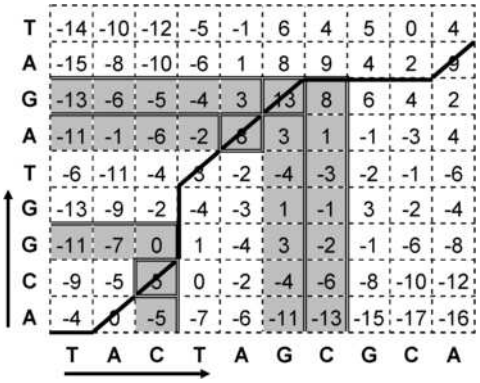

FIG. 1. *Gapped global alignment scores and the corresponding directed paths for two subsequences* $\mathbf{A}[1,10] = \mathsf{TACTAGCGCA}$ *and* $\mathbf{B}[1,9] = \mathsf{ACGGTAGAT}$*, drawn from the nucleotide alphabet* $\{\mathsf{A},\mathsf{C},\mathsf{G},\mathsf{T}\}$*. Figure 1 uses a nucleotide scoring matrix, where* $s(a,b) = 5$ *if* $a = b$ *and* -4 *otherwise, and the affine gap penalty* $w_g = 3 + 2g$*. The vertex* (i,j) *is in the northeast corner of the cell* (i,j)*, with the origin* $(0,0)$ *at the southwest corner of Figure 1. The cell* (i,j) *displays the global score* $S_{i,j}$*, calculated from (2.2). The optimal global path ending at the point* $(10,8)$*, for example, consists of* 12 *edges, in order:* 1 *east of length* $1,2$ *northeast,* 1 *north of length* $2,3$ *northeast,* 1 *east of length* 3*, and* 1 *northeast. The optimal global score* $S_{10,8} = -5+5+5-7+5+5+5-9+5 = 9$ *is the sum of the corresponding edges and represents the path of greatest weight starting at* $(0,0)$ *and ending at* $(10,8)$*. The corresponding optimal global alignment of the subsequences* $\mathbf{A}[1,10]$ *and* $\mathbf{B}[1,9]$ *is*

$$\mathsf{TAC{-}{-}TAGCGCA}$$
$$\mathsf{{-}ACGGTAG{-}{-}{-}A}.$$

The edge maxima are $M_1 = -4, M_2 = 0, M_3 = 5, M_4 = 1, M_5 = 3, M_6 = 8, M_7 = 13, M_8 = 9$*,* $M_9 = 6$*. The shading and the double lines indicate squares where a vertex (surrounded by double lines) generated an SALE* $\beta(k)$*. The SALE scores are* $M_{\beta(1)} = M_3 = 5, M_{\beta(2)} = M_6 = 8, M_{\beta(3)} = M_7 = 13$*; and the global maximum* M *for* $\mathbf{A}$ *and* $\mathbf{B}$ *is no less than* 13*, the largest global score shown.*

each $i = 1,2,\ldots,k$, the directed edge e_i comes out of vertex v_{i-1} and goes into vertex v_i. We say that the path π starts at v_0 and ends at v_k.

Denote finite subsequences of the sequence $\mathbf{A}$ by $\mathbf{A}[i,m] = A_i A_{i+1} \cdots A_m$. Every gapped alignment of the subsequences $\mathbf{A}[i,m]$ and $\mathbf{B}[j,n]$ corresponds to exactly one path that starts at $v_0 = (i-1, j-1)$ and ends at $v_k = (m,n)$ (see Figure 1). The alignment's score is the "path weight" $S_\pi := \sum_{i=1}^{k} s[e_i]$.

Define the "global score" $S_{i,j} := \max_\pi S_\pi$, where the maximum is taken over all paths π starting at $v_0 = (0,0)$ and ending at $v_k = (i,j)$. The paths π starting at v_0, ending at v_k, and having weight $S_\pi = S_{i,j}$ are "optimal global paths" and correspond to "optimal global alignments" between $\mathbf{A}[1,i]$ and $\mathbf{B}[1,j]$. Define the "edge maximum" $M_n := \max\{\max_{0\le i\le n} S_{i,n}, \max_{0\le j\le n} S_{n,j}\}$, and the "global maximum" $M := \sup_{n\ge 0} M_n$. (The single subscript in M_n indicates that the variate corresponds to a square $[0,n]\times[0,n]$, rather than a general rectangle $[0,m]\times[0,n]$.) Define the "strict ascending ladder epochs"

(SALEs) in the sequence (M_n): let $\beta(0) := 0$ and $\beta(k+1) := \min\{n > \beta(k) : M_n > M_{\beta(k)}\}$, where $\min \varnothing := \infty$. We call $M_{\beta(k)}$ the "kth SALE score."

Define also the "local score" $\tilde{S}_{i,j} := \max_\pi S_\pi$, where the maximum is taken over all paths π ending at $v_k = (i,j)$, regardless of their starting point. Define the "local maximum" $\tilde{M}_{m,n} := \max_{0\le i\le m, 0\le j\le n} \tilde{S}_{i,j}$. The paths π ending at $v_k = (i,j)$ with local score $S_\pi = \tilde{S}_{i,j} = \tilde{M}_{m,n}$ are "optimal local paths" corresponding to the "optimal local alignments" between subsequences of $\mathbf{A}[1,m]$ and $\mathbf{B}[1,n]$.

Now, the following "independent letters" model introduces randomness. Choose each letter in the sequence $\mathbf{A}$ and $\mathbf{B}$ randomly and independently from the alphabet $\mathfrak{L}$ according to fixed probability distributions $\{p_a : a \in \mathfrak{L}\}$ and $\{p'_b : b \in \mathfrak{L}\}$. (Although this article permits the distributions $\{p_a\}$ and $\{p'_b\}$ to be different, in applications they are usually the same.) Throughout the paper, the probability and expectation for the independent letters model are denoted by $\mathbb{P}$ and $\mathbb{E}$.

Let $\Gamma = \Gamma_{\mathbf{A},\mathbf{B}}$ denote the random alignment graph of the sequence-pair $(\mathbf{A},\mathbf{B})$. In the appropriate limit, if the alignment parameters are in the so-called "logarithmic phase" [6, 12] (i.e., if the optimal global alignment score of long random sequences has a negative score), the random local maximum $\tilde{M}_{m,n}$ follows an approximate Gumbel extreme value distribution with "scale parameter" λ and "pre-factor" K [1, 14],

$$\mathbb{P}(\tilde{M}_{m,n} > y) \approx 1 - \exp[-Kmn\exp(-\lambda y)]. \tag{2.1}$$

2.2. *The dynamic programming algorithm for global sequence alignment.* For affine gaps $w_g = \Delta_0 + \Delta_1 g$, the global score $S_{i,j}$ is calculated with the recursion

$$S_{i,j} = \max\{S_{i-1,j-1}, I_{i-1,j-1}, D_{i-1,j-1}\} + s(A_i, B_j), \tag{2.2}$$

where

$$I_{i,j} = \max\{S_{i,j-1} - \Delta_0 - \Delta_1, I_{i,j-1} - \Delta_1, D_{i,j-1} - \Delta_0 - \Delta_1\},$$

$D_{i,j} = \max\{S_{i-1,j} - \Delta_0 - \Delta_1, D_{i-1,j} - \Delta_1\}$ and boundary conditions $S_{0,0} = 0, I_{0,0} = D_{0,0} = -\infty, D_{g,0} = I_{0,g} = -\Delta_0 - \Delta_1 g, S_{g,0} = S_{0,g} = I_{g,0} = D_{0,g} = -\infty$ for $g > 0$ [15]. The three array names, S, I, and D, are mnemonics for "substitution," "insertion" and "deletion." If "Δ" denotes a gap character, the corresponding alignment letter-pairs $(a,b), (\Delta,b)$ and (a,Δ) correspond to the operations for editing sequence $\mathbf{A}$ into sequence $\mathbf{B}$ [30].

2.3. *Previous methods for estimating the BLAST p-value.* If $w_g \equiv \infty$ identically, so northward and eastward (gap) edges are disallowed in an optimal alignment path, a rigorous proof of (2.1) yields analytic formulas

for the Gumbel parameters λ and K [12]. For gapped local alignment, rigorous results are sparse, although some approximate analytical studies are extant [21, 22, 27, 29]. The prevailing approach therefore estimates λ and K from simulations [4, 31]. Because λ is an exponential rate, it dominates K's contribution to the BLAST p-value. Most studies therefore (including the present one) have focused on λ. (Note, however, some recent progress on the real-time estimation of K [26].) Typically, current applications require a 1–4% relative error in λ; 10–20%, in K [4]. The characteristics of the relevant sequence database determine the actual accuracies required, however, making approximations with controlled error and of arbitrary accuracy extremely desirable in practice.

Storey and Siegmund [29] approximate λ (with neither controlled errors nor arbitrary accuracy) as

$$\tilde{\lambda} \approx \lambda^* - 2(\mu^*)^{-1}\Lambda e^{-\lambda^*\Delta_0}/(e^{\lambda^*\Delta_1} - 1), \tag{2.3}$$

where $\sum_{(a,b)} p_a p'_b \exp[\lambda^* s(a,b)] = 1$ [so λ^* is the so-called "ungapped lambda," for $\Delta(g) \equiv \infty$] and $\mu^* := \sum_{(a,b)} s(a,b) p_a p'_b \exp[\lambda^* s(a,b)]$. In (2.3), Λ is an upper bound for an infinite sequence of constants defined in terms of gap lengths in a random alignment.

Many other studies have used local alignment simulations to estimate BLAST p-values, for example, Chan [9] used importance sampling and a mixture distribution. Some rigorous results [28] are also extant for the so-called "island method" [31, 32], which yields maximum likelihood estimates of λ and K from a Poisson process associated with local alignments exceeding a threshold score [4, 23].

Large deviations arguments [6, 35] support the common belief that global alignment can estimate λ for local alignment through the equation $\lambda = -\lim_{y\to\infty} y^{-1} \ln \mathbb{P}\{M \geq y\}$. For a fixed error, global alignment typically requires less computational effort than local alignment. For example, one early study [34] used importance sampling based on trial distributions $\mathbb{Q}$ from a hidden Markov model.

The study demonstrated that the global alignment equation $\mathbb{E}[\exp(\lambda S_{n,n})] = 1$ estimated λ with only $O(n^{-1})$ error [7]. (Recall that "$\mathbb{E}$" denotes the expectation corresponding to the random letters model.) The equation $\mathbb{E}[\exp(\lambda M_m)] = \mathbb{E}[\exp(\lambda M_n)]$ $(m \neq n)$, suggested by heuristic modeling with Markov additive processes (MAPs) [5, 10], improved the error substantially, to $O(\varepsilon^n)$ [24].

The next subsection shows how the MAP heuristic can improve the efficiency of importance sampling even further, with its renewal structure. The next subsection gives the relevant parts of the MAP heuristic.

2.4. *The Markov additive process heuristic.* The rigorous theory of MAPs appears elsewhere [5, 10]. Because the MAP heuristics given below parallel a previous publication [24], we present only informal essentials.

Consider a finite Markov-chain state-space $\mathfrak{J}$, containing $\#\mathfrak{J}$ elements. Without loss of generality, $\mathfrak{J} = \{1, \ldots, \#\mathfrak{J}\}$. Until further notice, all vectors are row vectors of dimension $\#\mathfrak{J}$; all matrices, of dimension $(\#\mathfrak{J}) \times (\#\mathfrak{J})$. A MAP can be defined in terms of a time-homogenous Markov chain (MC) $(J_n \in \mathfrak{J} : n = 0, 1, \ldots)$ and a $(\#\mathfrak{J}) \times (\#\mathfrak{J})$ matrix of real random variates $\|Z_{i,j}\|$. Let the MC have transition matrix $\mathbf{P} = \|p_{i,j}\|$, so $p_{i,j} = \mathbb{P}(J_n = j | J_{n-1} = i)$. Let the stationary distribution of the MC be $\boldsymbol{\pi}$, assumed strictly positive and satisfying both $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$ and $\boldsymbol{\pi}\mathbf{1}^t = 1$, where $\mathbf{1}^t$ denotes the $(\#\mathfrak{J}) \times 1$ column vector whose elements are all 1.

As usual, let $\mathbb{P}_{\boldsymbol{\gamma}}$ and $\mathbb{E}_{\boldsymbol{\gamma}}$ be the probability measure and expectation corresponding to an initial state J_0 with distribution $\boldsymbol{\gamma}$; $\mathbb{P}_i$ and $\mathbb{E}_i$, to an initial state $J_0 = i$; and $\mathbb{P}_{\boldsymbol{\pi}}$ and $\mathbb{E}_{\boldsymbol{\pi}}$, to an initial state in the equilibrium distribution $\boldsymbol{\pi}$.

Run the MC (J_n), and take its succession of states as given. Consider the following sequence $(Y_n \in \mathbb{R} : n = 0, 1, \ldots)$ of random variates. Define $Y_0 := 0$. For $n = 1, 2, \ldots$, let the (Y_n) be conditionally independent, with distributions determined by the transition $J_{n-1} \to J_n$ of the Markov chain as follows. If $J_{n-1} = i$ and $J_n = j$, the value of Y_n is chosen randomly from the distribution of $Z_{i,j}$. (Thus, if $J_{m-1} = J_{n-1} = i$ and $J_m = J_n = j, Y_m$ and Y_n share the distribution of $Z_{i,j}$, although independence permits randomness to give them different values.)

The random variates of central interest are the sums $T_n = \sum_{m=0}^{n} Y_m$ $(n = 0, 1, \ldots)$ and the maximum $M := \max_{n \geq 0} T_n$. To exclude trivial distributions for M (i.e., $M = 0$ a.s. and $M = \infty$ a.s.), make two assumptions: (1) $\mathbb{E}_{\boldsymbol{\pi}} Y_1 < 0$; and (2) there is some m and state i such that

$$\mathbb{P}_i\{\min\{T_k : k = 1, \ldots, m\} > 0; J_m = i, J_j \neq i \text{ for } j = 1, \ldots, m-1\} > 0. \tag{2.4}$$

Consider the sequence (T_n), its SALEs $\beta(0) := 0$ and $\beta(k+1) := \min\{n > \beta(k) : T_n > T_{\beta(k)}\}$, and its SALE scores $T_{\beta(k)}$. For brevity, let $\beta := \beta(1)$. Note that $M = T_{\beta(k)}$ for some $k \in \{0, 1, \ldots\}$. In a MAP, $(J_{\beta(k)}, T_{\beta(k)})$ forms a defective Markov renewal process.

Now, define the matrix $\mathbf{L}_\theta := \|\mathbb{E}_i[\exp(\theta T_\beta); J_\beta = j, \beta < \infty]\|$. The Perron–Frobenius theorem [5], page 25, shows that $\mathbf{L}_\theta$ has a strictly dominant eigenvalue $\rho(\theta) > 0$ [i.e., $\rho(\theta)$ is the unique eigenvalue of greatest absolute value]. Moreover, $\rho(\theta)$ is a convex function [19], and because $\mathbf{L}_0$ is substochastic, $\rho(0) < 1$. The two assumptions above (2.4) ensure that $M := \max_{n \geq 0} T_n$ has a nontrivial distribution and that $\rho(\lambda) = 1$ for some unique $\lambda > 0$.

The notation intentionally suggests a heuristic analogy between MAPs and global alignment. Identify the Markov chain states J_n in the MAP with

the rectangle $[0,n]\times[0,n]$ of $\Gamma_{\mathbf{A},\mathbf{B}}$, and identify the sum T_n in the MAP with the edge maximum M_n in global alignment. In the following, therefore, the identification leads to M_n replacing T_n in the MAP formulas. In particular, the MAP heuristic identifies the Gumbel scale parameter in (2.1) with the root $\lambda>0$ of the equation $\rho(\lambda)=1$. Although the heuristic analogy between MAPs and global alignment is in no way precise or rigorous, it has produced useful results [24].

The details of why the MAP heuristic works so well are presently obscure, although some additional motivation appears in an heuristic calculation related to λ [8]. The calculation takes the limit of nested successively wider semi-infinite strips, each strip having constant width and propagating itself northeastward in the alignment graph $\Gamma_{\mathbf{A},\mathbf{B}}$. The successive northeast boundaries of the propagation are states in an ergodic MC. MAPs therefore might rigorously justify the heuristic calculation.

3. Methods.

3.1. *A novel equation for λ.* From the definition of $\mathbf{L}_\theta$ in a MAP, if the Markov chain $\{J_n\}$ starts in a state J_0 with distribution $\boldsymbol{\gamma}$ (with M_n replacing T_n in the MAP formulas), matrix algebra applied to the concatenation of SALEs in a MAP yields

$$\mathbb{E}_{\boldsymbol{\gamma}}[\exp(\theta M_{\beta(k)});\beta(k)<\infty]=\boldsymbol{\gamma}(\mathbf{L}_\theta)^k\mathbf{1}^t. \tag{3.1}$$

For a MAP, equation (3.1) is exact; but for global alignment, it has no literal meaning. Equation (3.1) has some consequences for the limit $k\to\infty$, and we speculate that the consequences hold, even for global alignment. [Note: although the sequence $(\beta(k))$ is a.s. finite, the limits $k\to\infty$ below involve no contradiction or approximation, because they are not a.s. limits.]

Define $K_k(\theta):=\ln\{\mathbb{E}_{\boldsymbol{\gamma}}[\exp(\theta M_{\beta(k)});\beta(k)<\infty]\}$. In (3.1), a spectral (eigenvalue) decomposition of the matrix $\mathbf{L}_\theta$ [25] shows that

$$K_k(\theta)=k\ln\{\rho(\theta)\}+c_0+O(\varepsilon^k), \tag{3.2}$$

where $0\le\varepsilon<1$ is determined by the magnitude of the subdominant eigenvalue of $\mathbf{L}_\theta$, and c_0 is a constant independent of θ and k.

For $k'-k>0$ fixed, we can accelerate the convergence in (3.2) as $k\to\infty$ by differencing

$$K_{k'}(\theta)-K_k(\theta)=(k'-k)\ln\{\rho(\theta)\}+O(\varepsilon^k). \tag{3.3}$$

Let $\lambda_{k',k}$ denote the root of (3.3) after dropping the error term $O(\varepsilon^k)$. Because $\rho(\lambda)=1$, Taylor approximation around λ yields $\ln\{\rho(\lambda_{k',k})\}\approx\rho'(\lambda)(\lambda_{k',k}-\lambda)$, so (3.3) becomes

$$(k'-k)\rho'(\lambda)(\lambda_{k',k}-\lambda)=O(\varepsilon^k), \tag{3.4}$$

that is, with $k' - k$ fixed, $\lambda_{k',k}$ converges geometrically to λ as the SALE index $k \to \infty$.

The initial state $\boldsymbol{\gamma}$ of global alignment has a deterministic distribution, namely the origin $(0, 0)$. Equation (3.3) for $\theta = \lambda$ therefore becomes

$$\mathbb{E}[\exp(\lambda M_{\beta(k')}); \beta(k') < \infty] = \mathbb{E}[\exp(\lambda M_{\beta(k)}); \beta(k) < \infty] \tag{3.5}$$

after dropping the geometric error $O(\varepsilon^k)$. Let $\hat{\lambda}_{k',k}$ be the root of (3.5).

3.2. *The trial distribution for importance sampling.* In (3.5), crude Monte Carlo simulation generating random sequence-pairs with the identical letters model $\mathbb{P}$ is inefficient for the following reason. When practical alignment scoring systems are used, $\mathbb{P}\{\beta(k) < \infty\} < 1$ for $k \geq 1$. For, example, the BLAST defaults (scoring matrix BLOSUM62, gap penalty $w_g = 11 + g$, and Robinson–Robinson letter frequencies), $\mathbb{P}\{\beta(4) < \infty\} \approx 0.047$, so only about 1 in 20 crude Monte Carlo simulations generate a fourth ladder point. Empirically in our importance sampling, however, Gumbel parameter estimation seemed most efficient when the stopping time corresponded to $\beta(4)$ (see below).

Importance sampling requires a trial distribution to determine $\hat{\lambda}_{k',k}$ from (3.5). By editing one sequence into another, a Markov chain model borrowed directly from a previous study [34] generates random sequence alignments, as follows.

Consider a Markov state space consisting of the set of alignment letter-pairs $\bar{\mathfrak{L}}^2$, where $\bar{\mathfrak{L}} := \mathfrak{L} \cup \{\Delta\}$, "$\Delta$" being a character representing gaps. The ordered pair (Δ, Δ) has probability 0, so a succession of Markov states corresponds to a global sequence alignment (see Figure 1), that is, to a path in the alignment graph $\Gamma_{\mathbf{A},\mathbf{B}}$. Ordered pairs other than (Δ, Δ) fall into three sets, corresponding to edit operations following (2.2): $S := \mathfrak{L} \times \mathfrak{L}$ [substitution, a bioinformatics term implicitly including identical letter-pairs (a, a)], $I := \{\Delta\} \times \mathfrak{L}$ (insertion); and $D := \mathfrak{L} \times \{\Delta\}$ (deletion). The sets S, I and D form "atoms" of the MC [13], page 203, as follows. (By definition, each atom of a MC is a set of all states with identical outgoing transition probabilities.)

From the set S, the transition probability to (a, b) is $t_{S,S} q_{a,b}$; to $(\Delta, b), t_{S,I} p'_b$; and to $(a, \Delta), t_{S,D} p_a$. From the set I, the transition probability to (a, b) is $t_{I,S} q_{a,b}$; to $(\Delta, b), t_{I,I} p'_b$; and to $(a, \Delta), t_{I,D} p_a$. From the set D, the transition probability to (a, b) is $t_{D,S} q_{a,b}$; to $(\Delta, b), t_{D,I} p'_b$; and to $(a, \Delta), t_{D,D} p_a$. Transition probabilities sum to 1, so the following restrictions apply: $\sum_{a,b \in \mathfrak{L}} q_{a,b} = 1, \sum_{b \in \mathfrak{L}} p'_b = 1, \sum_{a \in \mathfrak{L}} p_a = 1, t_{S,S} + t_{S,I} + t_{S,D} = 1$ (transit from the substitution atom), $t_{D,D} + t_{D,S} + t_{D,I} = 1$ (transit from the deletion atom) and $t_{I,I} + t_{I,S} + t_{I,D} = 1$ (transit from the insertion atom). Usually in practice, the term $t_{I,D} = 0$, to disallow insertions following a deletion. Our formulas retain the term, to exploit the resulting symmetry later.

In the terminology of hidden Markov models, S, I, D are hidden Markov states. $t_{i,j}$ for $i, j \in \{S, I, D\}$ are transition probabilities and $q_{a,b}, p'_b, p_a$ for $a, b \in \mathfrak{L}$ are emission probabilities from the state S, I, D, respectively.

As described elsewhere [34], numerical values for the Markov probabilities can be determined from the scores $s(a, b)$ and the gap penalty w_g. Note that the values are selected for statistical efficiency, although many other values also yield unbiased estimates for λ in the appropriate limit.

3.3. *Importance sampling weights and stopping times.* To establish notation, and to make connections to the Appendix and its mapping theorem, note that the MC above can be supported on a probability space $(\Omega, \mathsf{F}, \mathbb{Q})$, where each $\omega = (\pi, \mathbf{A}, \mathbf{B}) \in \Omega$ is an ordered triple. Here, π is an infinite path starting at the origin in the alignment graph $\Gamma_{\mathbf{A},\mathbf{B}}$; F is the set generated by cylinder sets in Ω (here, cylinder sets essentially consist of some finite path and the corresponding pair of subsequences); and $\mathbb{Q}$ is the MC probability distribution described above, started at the atom S, with expectation operator $\mathbb{E}_{\mathbb{Q}}$.

Let N be any stopping time for the sequence $(M_n : n = 0, 1, \ldots)$ of edge maxima for $\Gamma_{\mathbf{A},\mathbf{B}}$ (i.e., the sequence $\{M_0, \ldots, M_n\}$ determines whether $N \leq n$ or not). Because M_n is determined by $(\mathbf{A}[1, n], \mathbf{B}[1, n])$, N is also a stopping time for the sequence $\{(\mathbf{A}[1, n], \mathbf{B}[1, n]) : n = 0, 1, \ldots\}$. The stopping time of main interest here is $N = \beta(k)$, the kth ladder index of (M_n), where $k \geq 1$ is arbitrary. (As further motivation for the mapping theorem in the Appendix, other stopping times of possible interest include, for example, $N = n$, a fixed epoch [7], and $N = \beta(K_y)$, where $\beta(K_y) = \inf\{n : M_n \geq y\}$ is the index of first ladder-score outside the interval $(0, y)$.)

To use the mapping theorem, introduce the probability space $(\Omega'', \mathsf{F}'', \mathbb{P})$, where each $\omega'' = (\mathbf{A}, \mathbf{B}) \in \Omega''$ is an ordered pair. Here, $\mathbf{A}$ and $\mathbf{B}$ are sequences, F'' is the set generated by all cylinder sets in Ω'' (i.e., sets corresponding to pairs of finite subsequences) and $\mathbb{P}(A'') = \prod_{k=1}^{i} p_{A_k} \prod_{k=1}^{j} p'_{B_k}$, if the cylinder set A'' corresponds to the subsequence pair $(\mathbf{A}[1, i], \mathbf{B}[1, j])$. Given N, the theory of stopping times [5], page 414, can be used to construct a discrete probability space $(\Omega', \mathsf{F}', \mathbb{P})$, where each event $\omega' \in \Omega'$ is a finite-sequence pair $\omega' = (\mathbf{A}[1, N], \mathbf{B}[1, N])$, F' is the set of all subsets of Ω' and $\mathbb{P}(\omega') = \prod_{k=1}^{N(\omega')} p_{A_k} \prod_{k=1}^{N(\omega')} p'_{B_k}$.

Let $\mathsf{I}_{m,n} := \{(i, j) : i = m, j \geq n\}$ and $\mathsf{D}_{m,n} := \{(i, j) : i \geq m, j = n\}$. Define the function $f : \omega \mapsto \omega'$, where $\omega = (\pi, \mathbf{A}, \mathbf{B})$ and $\omega' = (\omega'_{\mathbf{A}}, \omega'_{\mathbf{B}}) := (\mathbf{A}[1, N], \mathbf{B}[1, N])$. Then, $\omega \in f^{-1}(\omega')$, if and only if the path π hits the set $\mathsf{I}_{N,N} \cup \mathsf{D}_{N,N}$ at (i, j), so that $\mathbf{A}[1, N] = \omega'_{\mathbf{A}}$ and $\mathbf{B}[1, N] = \omega'_{\mathbf{B}}$ (see Figure 2).

Empirically, our simulations satisfied $\mathbb{Q}\{\beta(k) < \infty\} = 1$, and we speculate that our application therefore satisfies the hypothesis $\mathbb{Q}H = 1$ of the Appendix. According to the Appendix, the reciprocal importance sampling

weight $1/W(\omega) = \sum_{\omega_0 \in f^{-1}\{f(\omega)\}} \mathbb{Q}(\omega_0)/\mathbb{P}f(\omega)$ depends on the sum over all possible Markov chain realizations $\omega_0 \in f^{-1}(\omega')$. Dynamic programming computes the sum efficiently, as follows.

Let the "transition" T represent any element of $\{S, I, D\}$ [substitution (a_i, b_j), insertion (Δ, b_j), or deletion (a_i, Δ)]. Fix any particular pair $(\mathbf{A}, \mathbf{B})$ of infinite sequences, which fixes $N = \beta(k)$. To set up a recursion for dynamic programming, consider the following set of events $\mathsf{E}^T_{i,j}$, defined for $T \in \{S, I, D\}$ and $\min\{i, j\} \le N$, and illustrated in Figure 2. Let $\mathsf{E}^T_{i,j}$ be the event consisting of all ω yielding a path π whose final transition is T and which corresponds to the subsequences: (1) $\mathbf{A}[1, i]$ and $\mathbf{B}[1, j]$ for $0 \le i, j \le N$; (2) $\mathbf{A}[1, i]$ and $\mathbf{B}[1, N]$ for $0 \le N = j \le i$; and (3) $\mathbf{A}[1, N]$ and $\mathbf{B}[1, j]$ for $0 \le N = i \le j$. Define $Q^T_{i,j} := \mathbb{Q}(\mathsf{E}^T_{i,j})$ and $Q_{i,j} := Q^S_{i,j} + Q^I_{i,j} + Q^D_{i,j}$.

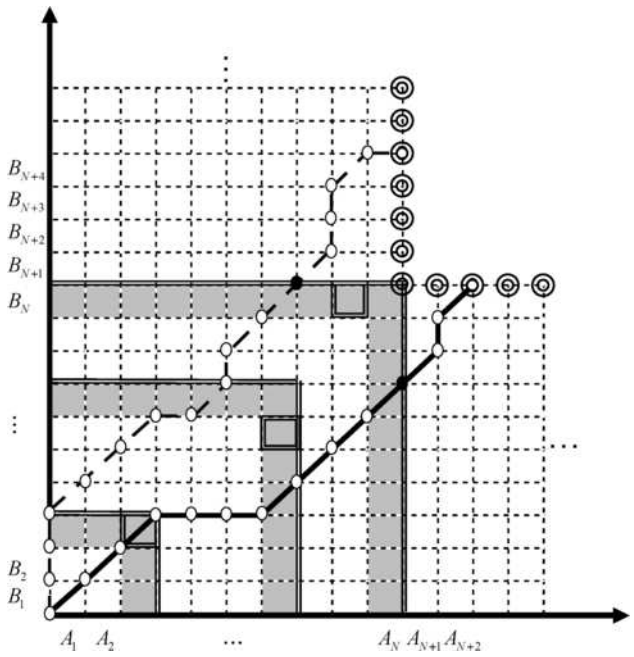

FIG. 2. *Two examples of alignment path π generated by a Markov chain. As in Figure 1, the shading and the double lines indicate squares where a vertex (surrounded by double lines) generated an SALE. The SALEs determine the stopping time $N = \beta(3)$. In Figure 2, the first SALE is determined by the score at the vertex $(3, 3)$; the second SALE, the vertex $(7, 6)$; the third SALE, the vertex $(9, 10)$. Therefore, $N = \beta(3) = 10$. The vertical ray $\mathsf{I}_{N,N}$ and the horizontal ray $\mathsf{D}_{N,N}$ are indicated by double circles. The lower path π (solid line) ends at $(N + 2, N)$ with a final transition to S; the upper path π (long-dashed line), at $(N, N + 4)$ with a final transition to D. The closed vertices indicate intersection with the square corresponding to $\omega' = (\omega'_{\mathbf{A}}, \omega'_{\mathbf{B}}) = (\mathbf{A}[1, N], \mathbf{B}[1, N])$.*

(Note: in the following, $T \in \{S, I, D\}$ is always a superscript, never an exponent.)

For brevity, let $\tilde{q}_{i,j} = q_{A_i,B_j}$ for $0 \le i, j \le N$; $\tilde{q}_{i,j} = \sum_{(a \in \mathfrak{L})} q_{a,B_j}$ for $0 \le j \le N < i$; $\tilde{q}_{i,j} = \sum_{(b \in \mathfrak{L})} q_{A_i,b}$ for $0 \le i \le N < j$; and $\tilde{q}_{i,j} = 1$ otherwise. Let $\tilde{p}'_j = p'_{B_j}$ for $0 \le j \le N$; and 1 otherwise. Finally, Let $\tilde{p}_i = p_{A_i}$ for $0 \le i \le N$; and 1 otherwise. Because every path into the vertex (i, j) comes from one of three vertices, each corresponding to a different transition $T \in \{S, I, D\}$,

$$\begin{aligned} Q^S_{i,j} &= \tilde{q}_{i,j}(t_{S,S}Q^S_{i-1,j-1} + t_{I,S}Q^I_{i-1,j-1} + t_{D,S}Q^D_{i-1,j-1}), \\ Q^I_{i,j} &= \tilde{p}'_j(t_{S,I}Q^S_{i,j-1} + t_{I,I}Q^I_{i,j-1} + t_{D,I}Q^D_{i,j-1}), \\ Q^D_{i,j} &= \tilde{p}_i(t_{S,D}Q^S_{i-1,j} + t_{I,D}Q^I_{i-1,j} + t_{D,D}Q^D_{i-1,j}) \end{aligned} \tag{3.6}$$

with boundary conditions $Q^S_{0,0} = 1, Q^I_{0,0} = Q^D_{0,0} = 0, Q^S_{g,0} = Q^S_{0,g} = Q^I_{g,0} = Q^D_{0,g} = 0, Q^I_{0,g} = p'_{B_1} \cdots p'_{B_g} t_{S,I}(t_{I,I})^{g-1}$ and $Q^D_{g,0} = p_{A_1} \cdots p_{A_g} t_{S,D} \times (t_{D,D})^{g-1}$ $(g > 0)$.

Recall that $\omega = (\pi, \mathbf{A}, \mathbf{B}) \in f^{-1}(\omega')$, if and only if the path π hits the set $\mathsf{I}_{N,N} \cup \mathsf{D}_{N,N}$ at (i, j), so that $\mathbf{A}[1, N] = \omega'_{\mathbf{A}}$ and $\mathbf{B}[1, N] = \omega'_{\mathbf{B}}$. Thus,

$$\sum_{\omega \in f^{-1}(\omega')} \mathbb{Q}(\omega) = -Q^S_{N,N} + \sum_{j=N}^{\infty} (Q^S_{N,j} + Q^D_{N,j}) + \sum_{i=N}^{\infty} (Q^S_{i,N} + Q^I_{i,N}). \tag{3.7}$$

To turn (3.6) into a recursion for importance sampling weights, define $P_i := p_{A_1} \cdots p_{A_{\min\{i,N\}}} = \tilde{p}_1 \cdots \tilde{p}_i$ and $P'_j := p'_{B_1} \cdots p'_{B_{\min\{j,N\}}} = \tilde{p}'_1 \cdots \tilde{p}'_j$, and let $W^T_{i,j} := Q^T_{i,j}/(P_i P'_j)$ $(T \in \{S, I, D\})$. Let $r_{i,j} = \tilde{q}_{i,j}/(\tilde{p}_i \tilde{p}'_j)$. For future reference, define $r_{\bullet,j} := r_{i,j}$ for $0 \le j \le N < i$ and $r_{i,\bullet} := r_{i,j}$ for $0 \le i \le N < j$. Note that $r_{\bullet,j}$ is independent of i, and $r_{i,\bullet}$ is independent of j. Equation (3.6) yields

$$\begin{aligned} W^S_{i,j} &= r_{i,j}(t_{S,S}W^S_{i-1,j-1} + t_{I,S}W^I_{i-1,j-1} + t_{D,S}W^D_{i-1,j-1}), \\ W^I_{i,j} &= t_{S,I}W^S_{i,j-1} + t_{I,I}W^I_{i,j-1} + t_{D,I}W^D_{i,j-1}, \\ W^D_{i,j} &= t_{S,D}W^S_{i-1,j} + t_{I,D}W^I_{i-1,j} + t_{D,D}W^D_{i-1,j} \end{aligned} \tag{3.8}$$

with boundary conditions $W^S_{0,0} = 1, W^I_{0,0} = W^D_{0,0} = 0, W^S_{g,0} = W^S_{0,g} = W^I_{g,0} = W^D_{g,0} = 0, W^I_{0,g} = t_{S,I}(t_{I,I})^{g-1}$ and $W^D_{g,0} = t_{S,D}(t_{D,D})^{g-1}$ $(g > 0)$. Because of (3.7), the importance sampling weight $W := W(\omega)$ satisfies

$$\begin{aligned} \frac{1}{W} &= \frac{\sum_{\omega_0 \in f^{-1}\{f(\omega)\}} \mathbb{Q}(\omega_0)}{\mathbb{P}f(\omega)} \\ &= -W^S_{N,N} + \sum_{j=N}^{\infty} (W^S_{N,j} + W^D_{N,j}) + \sum_{i=N}^{\infty} (W^S_{i,N} + W^I_{i,N}). \end{aligned} \tag{3.9}$$

Because $r_{i,j} = r_{\bullet,j}$ $(0 \le j \le N < i)$ and $r_{i,j} = r_{i,\bullet}$ $(0 \le i \le N < j)$, only a finite number of recursions are needed to compute the infinite sums in (3.9), as follows. For $T \in \{S, I, D\}$, define $\tilde{U}_i^T := U_{i,N}^T$, where $U_{m,n}^T := \sum_{j=n}^{\infty} W_{m,j}^T$. Likewise, define $\tilde{V}_j^T := V_{N,j}^T$, where $V_{m,n}^T := \sum_{i=m}^{\infty} W_{i,n}^T$. Equation (3.9) becomes

$$\frac{1}{W} = -W_{N,N}^S + \tilde{U}_N^S + \tilde{U}_N^D + \tilde{V}_N^S + \tilde{V}_N^I. \tag{3.10}$$

Note that $U_{i,j-1}^T - U_{i,j}^T = W_{i,j-1}^T$. To determine $\tilde{U}_N^T$, summation of (3.8) for $0 \le i \le N < j$ yields

$$\begin{aligned} U_{i,j}^S &= r_{i,\bullet}(t_{S,S}U_{i-1,j-1}^S + t_{I,S}U_{i-1,j-1}^I + t_{D,S}U_{i-1,j-1}^D) \\ &= U_{i,j-1}^S - W_{i,j-1}^S, \\ U_{i,j}^I &= t_{S,I}U_{i,j-1}^S + t_{I,I}U_{i,j-1}^I + t_{D,I}U_{i,j-1}^D = U_{i,j-1}^I - W_{i,j-1}^I, \\ U_{i,j}^D &= t_{S,D}U_{i-1,j}^S + t_{I,D}U_{i-1,j}^I + t_{D,D}U_{i-1,j}^D. \end{aligned} \tag{3.11}$$

Elimination of $U_{i,j}^T$ for $j = N+1$ and $i = 1, \ldots, N$ in the first two equations yields

$$\begin{aligned} U_{i,N}^S &= r_{i,\bullet}(t_{S,S}U_{i-1,N}^S + t_{I,S}U_{i-1,N}^I + t_{D,S}U_{i-1,N}^D) + W_{i,N}^S, \\ U_{i,N}^I &= t_{S,I}U_{i,N}^S + t_{I,I}U_{i,N}^I + t_{D,I}U_{i,N}^D + W_{i,N}^I, \\ U_{i,N}^D &= t_{S,D}U_{i-1,N}^S + t_{I,D}U_{i-1,N}^I + t_{D,D}U_{i-1,N}^D, \end{aligned} \tag{3.12}$$

that is,

$$\begin{aligned} \tilde{U}_i^S &= r_{i,\bullet}(t_{S,S}\tilde{U}_{i-1}^S + t_{I,S}\tilde{U}_{i-1}^I + t_{D,S}\tilde{U}_{i-1}^D) + W_{i,N}^S, \\ \tilde{U}_i^I &= (1 - t_{I,I})^{-1}(t_{S,I}\tilde{U}_i^S + t_{D,I}\tilde{U}_i^D + W_{i,N}^I), \\ \tilde{U}_i^D &= t_{S,D}\tilde{U}_{i-1}^S + t_{I,D}\tilde{U}_{i-1}^I + t_{D,D}\tilde{U}_{i-1}^D \end{aligned} \tag{3.13}$$

with initial values $\tilde{U}_0^S = \tilde{U}_0^D = 0$ and $\tilde{U}_0^I = (1 - t_{I,I})^{-1}W_{0,N}^I = (1 - t_{I,I})^{-1} \times t_{S,I}(t_{I,I})^{N-1}$. Compute (3.13) recursively for $i = 1, \ldots, N$.

Similarly, reflect through $i = j$ to derive

$$\begin{aligned} \tilde{V}_j^S &= r_{\bullet,j}(t_{S,S}\tilde{V}_{j-1}^S + t_{D,S}\tilde{V}_{j-1}^D + t_{I,S}\tilde{V}_{j-1}^I) + W_{N,j}^S, \\ \tilde{V}_j^I &= t_{S,I}\tilde{V}_{j-1}^S + t_{D,I}\tilde{V}_{j-1}^D + t_{I,I}\tilde{V}_{j-1}^I, \\ \tilde{V}_j^D &= (1 - t_{D,D})^{-1}(t_{S,D}\tilde{V}_j^S + t_{I,D}\tilde{V}_j^I + W_{N,j}^D) \end{aligned} \tag{3.14}$$

with initial values $\tilde{V}_0^S = \tilde{V}_0^I = 0$ and $\tilde{V}_0^D = (1 - t_{D,D})^{-1}W_{N,0}^D = (1 - t_{D,D})^{-1} \times t_{S,D}(t_{D,D})^{N-1}$. Iterate (3.14) for $j = 1, \ldots, N$. Substitute the results for $\tilde{U}_N^S, \tilde{U}_N^D, \tilde{V}_N^S$, and $\tilde{V}_N^I$ into (3.10) to compute W.

3.4. *Error estimates for* $\hat{\lambda}_{k',k}$. Denote the indicator of an event A by $\mathbb{I}A$, that is, $\mathbb{I}A = 1$ if A occurs and 0 otherwise. For a realization ω in the simulation, define

$$\begin{aligned} h_{k,k'}(\theta) &:= h_{k,k'}(\theta;\omega) \\ &:= \exp(\theta M_{\beta(k')})\mathbb{I}[\beta(k') < \infty] - \exp(\theta M_{\beta(k)})\mathbb{I}[\beta(k) < \infty] \end{aligned} \tag{3.15}$$

and let $h'_{k,k'}$ be its derivative with respect to θ.

Given samples ω_i $(i = 1, \ldots, r)$ from the trial distribution $\mathbb{Q}$, let $W = W(\omega_i)$ denote the corresponding importance sampling weights. Because $\hat{\lambda}_{k',k}$ is the M-estimator [17] of the root $\lambda_{k',k}$ of $\mathbb{E}h_{k,k'}(\lambda_{k',k}) = 0$, as $r \to \infty$, $\sqrt{r}(\hat{\lambda}_{k',k} - \lambda_{k',k})$ converges in distribution to the normal distribution with mean 0 and variance [17]

$$\frac{\mathbb{E}_{\mathbb{Q}}[h(\lambda_{k',k})W]^2}{\{\mathbb{E}_{\mathbb{Q}}[h'(\lambda_{k',k})W]\}^2} \approx \frac{r^{-1}\sum_1^r [h(\omega_i;\hat{\lambda}_{k',k})W(\omega_i)]^2}{\{r^{-1}\sum_1^r [h'(\omega_i;\hat{\lambda}_{k',k})W(\omega_i)]\}^2}. \tag{3.16}$$

4. Numerical study for Gumbel scale parameter. Table 1 gives our "best estimate" $\bar{\lambda}$ of the Gumbel scale parameter λ from (3.5) for each of the 5 options BLASTP gives users for the alignment scoring scheme. For every scheme, estimates $\hat{\lambda}$ derived from the first to fourth SALEs indicated that $\hat{\lambda}$ generally is biased above the true value λ, but that $\hat{\lambda}$ converged adequately by the fourth SALE. The best estimate $\bar{\lambda}$ (shown in Table 1) is the average of 200 independent estimates $\hat{\lambda}$, each computed within 1 sec from sequence-pairs simulated up to their fourth SALE. For BLOSUM 62 and gap penalty $w_g = 11 + g$, the average computation produced 1441 sequence-pairs up to their fourth SALE within 1 second. (For results relevant to the other publicly available scoring schemes, see Table 1.) The best estimates $\bar{\lambda}$ derived from (3.5) were within the error of the BLASTP values for λ.

Despite having the variance formula in (3.16) in hand, we elected to estimate the standard error $\hat{s}_\lambda$ directly from the 200 independent estimates $\hat{\lambda}$. Figure 3 plots the relative error $\hat{s}_\lambda/\bar{\lambda}$ in each individual $\hat{\lambda}$ against the computation time, where $\hat{s}_\lambda$ is the standard error of $\hat{\lambda}$. It shows that for all 5 BLASTP online options, (3.5) easily computed $\hat{\lambda}$ to 1–4% accuracy within about 0.5 seconds.

5. Discussion. This article indicates that the scale parameter λ of the Gumbel distribution for local alignment of random sequences satisfies (3.5), an equation involving the strict ascending ladder-points (SALEs) from global alignment, at least approximately. For standard protein scoring systems, in fact, simulation error could account for most (if not all) of the observed differences between values of λ calculated from (3.5) and values calculated from

TABLE 1

Best estimates $\bar{\lambda}$ for the 5 BLASTP alignment scoring schemes. For each scheme, we generated 200 estimates $\hat{\lambda}$, each within a one-second computation time. The third column gives present estimates of λ used on the BLAST web page (Stephen Altschul: personal communication). The BLAST values are accurate to approximately ±1%. The fourth column gives the mean $\bar{\lambda}$ of our 200 estimates $\hat{\lambda}$; the fifth, the standard error of $\bar{\lambda}$, which can be multiplied by $\sqrt{200} \approx 14$ to give the standard error in each $\hat{\lambda}$. The sixth column gives the average number of sequence-pairs used to estimate each $\hat{\lambda}$. The total number of sequence-pairs used for $\bar{\lambda}$ is 200 times average number of sequence-pairs. The last column gives the average sequence length required for the fourth SALE used to estimate each $\hat{\lambda}$

Scoring matrix	**Gap penalty w_g**	**BLAST value**	**Best estimate $\bar{\lambda}$**	**Standard error of $\bar{\lambda}$**	**Average number of sequence-pairs**	**Average sequence length**
BLOSUM80	$10+g$	0.299	0.2998	0.0001	2865	15.85
BLOSUM62	$11+g$	0.267	0.2679	0.0002	1441	27.78
BLOSUM45	$14+2g$	0.195	0.1962	0.0003	789	39.23
PAM30	$9+g$	0.294	0.2956	0.0001	3593	9.20
PAM70	$10+g$	0.291	0.2922	0.0001	3397	11.49

extensive crude Monte Carlo simulations. (The values of λ from crude simulation have a standard error of about ±1%.) In SALE simulations, (3.5) estimated λ to 1–4% accuracy within 0.5 second, as required by BLAST database searches over the Web. The present study did not tune simulations much; it relied instead on methods specific to sequence alignment to improve estimation. Many general strategies for sequential importance sampling therefore remain available to speed simulation. Preliminary investigations estimating the other Gumbel parameter (the pre-factor K) with SALEs are encouraging, so online estimation of the entire Gumbel distribution for arbitrary scoring schemes appears imminent, and preliminary computer code is already in place.

APPENDIX: A GENERAL MAPPING THEOREM FOR IMPORTANCE SAMPLING

The following theorem describes an unusual type of Rao-Blackwellization [20]. Consider two probability spaces $(\Omega, \mathsf{F}, \mathbb{Q})$ and $(\Omega', \mathsf{F}', \mathbb{P})$, and a F/F'-measurable function $f : \Omega \mapsto \Omega'$ (i.e., $f^{-1}F' \in \mathsf{F}$ for every $F' \in \mathsf{F}'$). Note: f is explicitly permitted to be many-to-one. Let $\mathbb{P} << \mathbb{Q}f^{-1}$ on some set H' (i.e., $\mathbb{Q}f^{-1}G' = 0 \Rightarrow \mathbb{P}G' = 0$ for any set $G' \subseteq H'$), so the Radon–Nikodym derivative in the second line of (A.1) below exists. Let $H := f^{-1}H'$, so for every random variate X' on (Ω', F'),

$$\mathbb{E}[X'; H'] := \int_{\omega' \in H'} X'(\omega')\, d\mathbb{P}(\omega')$$

$$
\begin{aligned}
&= \int_{\omega' \in H'} X'(\omega')\, d\mathbb{P}(\omega') \int_{\omega \in f^{-1}(\omega')} \frac{d\mathbb{Q}(\omega)}{\int_{\omega_0 \in f^{-1}(\omega')} d\mathbb{Q}(\omega_0)} \\
&= \int_{\omega' \in H'} \int_{\omega \in f^{-1}(\omega')} X' f(\omega) \frac{d\mathbb{P} f(\omega)}{\int_{\omega_0 \in f^{-1}\{f(\omega)\}} d\mathbb{Q}(\omega_0)}\, d\mathbb{Q}(\omega) \\
&= \int_{\omega \in H} X' f(\omega) \frac{d\mathbb{P} f(\omega)}{\int_{\omega_0 \in f^{-1}\{f(\omega)\}} d\mathbb{Q}(\omega_0)}\, d\mathbb{Q}(\omega).
\end{aligned} \tag{A.1}
$$

Consider the application of (A.1) to importance sampling with target distribution $\mathbb{P}$ and trial distribution $\mathbb{Q}$. Assume $\mathbb{Q}H = 1$, so H supports $\mathbb{Q}$. In our application to global alignment, $H = [\beta(k) < \infty] \subset \Omega$ ("$\subset$" being strict inclusion), but we speculate $\mathbb{Q}H = 1$.

In Monte Carlo applications, a discrete sample space H is usually available. Accordingly, the following theorem replaces the integrals in (A.1) by sums.

The mapping theorem for importance sampling. *Let*

$$
\frac{1}{W(\omega)} := \frac{\sum_{\omega_0 \in f^{-1}\{f(\omega)\}} \mathbb{Q}(\omega_0)}{\mathbb{P} f(\omega)}. \tag{A.2}
$$

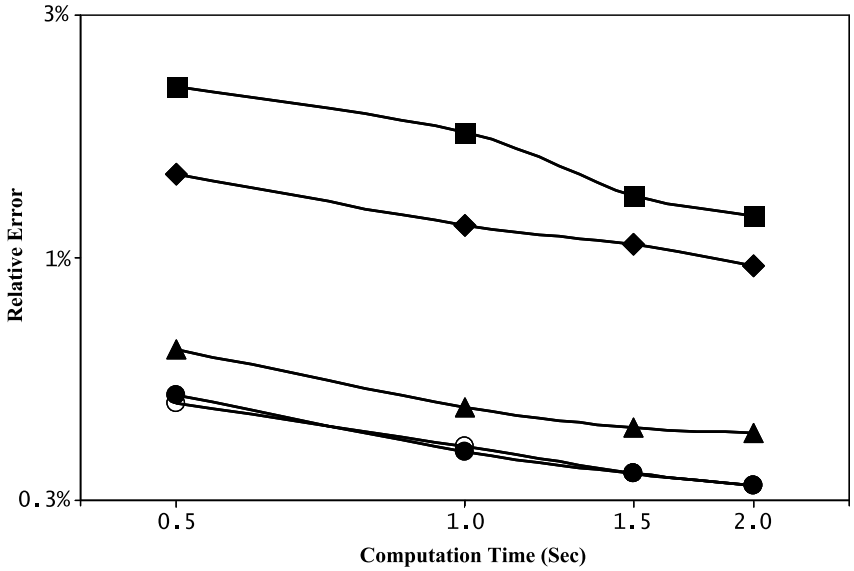

FIG. 3. *Plot of relative errors against computation time (sec). Both axes are in logarithmic scale. Computation time was measured on a 2.99 GHz Pentium® D CPU. Relative errors for BLOSUM45 with* $\Delta(g) = 14 + 2g$ *are shown by* ■*; BLOSUM62 with* $\Delta(g) = 11 + g$*, by* ♦*; BLOSUM80 with* $\Delta(g) = 10 + g$*, by* Q*; PAM70 with* $\Delta(g) = 10 + g$*, by* "*; PAM30 with* $\Delta(g) = 9 + g$*, by* !.

Under the above conditions, $r^{-1}\sum_{i=1}^{r}[X'f(\omega_i)W(\omega_i)] \to \mathbb{E}[X';H']$ *with probability* 1 *and in mean* (*with respect to* $\mathbb{Q}$), *as the number of realizations* $r \to \infty$.

The mapping theorem is an easy application of the law of large numbers to (A.1).

Acknowledgments. The authors Y. Park and S. Sheetlin contributed equally to the article. All authors would like to acknowledge helpful discussion with Dr. Nak-Kyeong Kim. This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

REFERENCES

[1] Aldous, D. (1989). *Probability Approximations via the Poisson Clumping Heuristic*, 1st ed. Springer, New York. MR0969362

[2] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool. *J. Molecular Biology* **215** 403–410.

[3] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25** 3389–3402.

[4] Altschul, S. F., Bundschuh, R., Olsen, R. and Hwa, T. (2001). The estimation of statistical parameters for local alignment score distributions. *Nucleic Acids Res.* **29** 351–361.

[5] Asmussen, S. (2003). *Applied Probability and Queues*. Springer, New York. MR1978607

[6] Arratia, R. and Waterman, M. S. (1994). A phase transition for the score in matching random sequences allowing deletions. *Ann. Appl. Probab.* **4** 200–225. MR1258181

[7] Bundschuh, R. (2002). Rapid significance estimation in local sequence alignment with gaps. *J. Comput. Biology* **9** 243–260.

[8] Bundschuh, R. (2002). Asymmetric exclusion process and extremal statistics of random sequences. *Phys. Rev. E* **65** 031911.

[9] Chan, H. P. (2003). Upper bounds and importance sampling of p-values for DNA and protein sequence alignments. *Bernoulli* **9** 183–199. MR1997026

[10] Cinlar, E. (1975). *Introduction to Stochastic Processes*. Prentice Hall, Upper Saddle River, NJ. MR0380912

[11] Dayhoff, M. O., Schwartz, R. M. and Orcutt, B. C. (1978). A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure* 345–352. National Biomedical Research Foundation, Silver Spring, MD.

[12] Dembo, A., Karlin, S. and Zeitouni, O. (1994). Limit distributions of maximal nonaligned two-sequence segmental score. *Ann. Probab.* **22** 2022–2039. MR1331214

[13] Djellout, H. and Guillin, A. (2001). Moderate deviations for Markov chains with atom. *Stochastic Process. Appl.* **95** 203–217. MR1854025

[14] Galombos, J. (1978). *The Asymptotic Theory of Extreme Order Statistics*, 1st ed. Wiley and Sons, New York.

[15] Gotoh, O. (1982). An improved algorithm for matching biological sequences. *J. Molecular Biology* **162** 705–708.

[16] Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **89** 10915–10919.

[17] Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35** 73–101. MR0161415

[18] Karlin, S. and Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. In *Proceedings of the National Academy of Sciences of the United States of America* **87** 2264–2268.

[19] Kingman, J. F. C. (1961). A convexity property of positive matrices. *Quart. J. Math. Oxford* **12** 283–284. MR0138632

[20] Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer, New York. MR1842342

[21] Mott, R. (1999). Local sequence alignments with monotonic gap penalties. *Bioinformatics* **15** 455–462.

[22] Mott, R. (2000). Accurate formula for p-values of gapped local sequence and profile alignments. *J. Molecular Biology* **300** 649–659.

[23] Olsen, R., Bundschuh, R. and Hwa, T. (1999). Rapid assessment of extremal statistics for gapped local alignment. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology* 211–222. AAAI Press, Menlo Park, CA.

[24] Park, Y., Sheetlin, S. and Spouge, J. L. (2005). Accelerated convergence and robust asymptotic regression of the Gumbel scale parameter for gapped sequence alignment. *J. Phys. A: Mathematical and General* **38** 97–108.

[25] Seneta, E. (1981). *Nonnegative Matrices and Markov Chain*. Springer, New York. MR0719544

[26] Sheetlin, S., Park, Y. and Spouge, J. L. (2005). The Gumbel pre-factor k for gapped local alignment can be estimated from simulations of global alignment. *Nucleic Acids Res.* **33** 4987–4994.

[27] Siegmund, D. and Yakir, B. (2000). Approximate p-values for local sequence alignments. *Ann. Statist.* **28** 657–680. MR1792782

[28] Spouge, J. L. (2004). Path reversal, islands, and the gapped alignment of random sequences. *J. Appl. Probab.* **41** 975–983. MR2122473

[29] Storey, J. D. and Siegmund, D. (2001). Approximate p-values for local sequence alignments: Numerical studies. *J. Comput. Biology* **8** 549–556.

[30] Waterman, M. S., Smith, T. F. and Beyer, W. A. (1976). Some biological sequence metrics. *Adv. in Math.* **20** 367–387. MR0408876

[31] Waterman, M. S. and Vingron, M. (1994). Rapid and accurate estimates of statistical significance for sequence data base searches. *Proc. Natl. Acad. Sci. USA* **91** 4625–4628.

[32] Waterman, M. S. and Vingron, M. (1994). Sequence comparison significance and Poisson approximation. *Statist. Sci.* **9** 367–381. MR1325433

[33] Yu, Y. K. and Altschul, S. F. (2005). The construction of amino acid substitution matrices for the comparison of proteins with nonstandard compositions. *Bioinformatics* **21** 902–911.

[34] Yu, Y. K. and Hwa, T. (2001). Statistical significance of probabilistic sequence alignment and related local hidden Markov models. *J. Comput. Biology* **8** 249–282.

[35] Zhang, Y. (1995). A limit theorem for matching random sequences allowing deletions. *Ann. Appl. Probab.* **5** 1236–1240. MR1384373

Y. Park
S. Sheetlin
J. L. Spouge
National Center
for Biotechnology Information
National Library of Medicine
National Institutes
of Health
8600 Rockville Pike
Bethesda, Maryland 20894
USA
E-mail: park@ncbi.nlm.nih.gov
sheetlin@ncbi.nlm.nih.gov
spouge@ncbi.nlm.nih.gov